Summary of “Vulnerabilities in Privacy-Preserving Record Linkage: The Threat of Dataset Extension Attacks”

Marcel Mildenberger

October 5, 2025

1 Motivation and Research Questions

The growing demand for cross-institutional data sharing in healthcare, official statistics, and scientific research has made Privacy-Preserving Record Linkage (PPRL) a critical enabling technology. PPRL enables the integration of datasets across organizational boundaries while ensuring that sensitive Personally Identifiable Information (PII) remains protected during probabilistic record linkage process. Privacy enhancing encodings such as Bloom Filter (BF) are widely adopted because they preserve similarity relationships between records while preventing exposure of raw PII during linkage. However, practical deployments increasingly face attackers equipped with new strategies. This thesis examines how attackers can exploit previously leaked matches, such as those obtained through a Graph Matching Attack (GMA), to compromise additional records that were previously unidentifiable. The central research questions are threefold: (i) to what extent can the capabilities of current state-of-the-art GMAs be extended through subsequent attacks, (ii) how effectively can Artificial Neural Network-based models learn the structural regularities of encoded identifiers, and (iii) which encoding strategies demonstrate resilience against such attacks under realistic conditions, and which remain particularly vulnerable.

The relevance of this analysis lies in the practical risks associated with the potential re-identification of encoded sensitive data such as health care data. The combination of graph-based and pattern-learning attacks introduces new and largely unexplored attack surfaces. While prior research has primarily focused on vulnerabilities in BF-based linkage and graph-matching techniques, little empirical evidence has addressed the feasibility of extending re-identification beyond this scope. This thesis closes that gap by proposing and evaluating the Dataset Extension Attack (DEA), a two-stage attack that employs neural models to infer plaintext n-grams from encoded representations. Through systematic evaluation across varying dataset overlaps and realistic conditions, this work provides an empirically grounded assessment of the privacy risks faced by current PPRL deployments.

2 Foundations of Privacy-Preserving Record Linkage

2.1 Threat Model and Similarity-Preserving Encodings

The thesis considers a classical PPRL scenario involving two data holders and a linkage unit. One data holder, Alice, maintains a database of encoded records D_e that she intends to link with another party without revealing any underlying PII. The linkage unit, Eve, performs the record linkage on behalf of both data holders but is modeled as an honest-but-curious attacker. She follows the prescribed protocol correctly, with full knowledge of the encoding parameters but without access to any secret values, while simultaneously attempting to infer and re-identify records in D_e.

To this end, Eve leverages an auxiliary plaintext dataset D_p that partially overlaps with Alice's database. This auxiliary dataset may originate from publicly available information, previously leaked data, or other external sources. Eve's objective is to maximize the number of records in D_e that she can re-identify by employing attacks such as the GMA and the proposed DEA.

The study focuses on the three most used encoding schemes in PPRL to assess their respective vulnerabilities. BF encoding maps n-grams into a fixed-length binary vector using multiple hash functions. Tabulation MinHash (TMH) employs tabulation-based MinHashing to generate fixed-length binary vectors with improved resistance to frequency-based attacks. Two-Step Hash (TSH) first encodes data using BFs and subsequently applies an additional hashing layer to produce integer vectors, trying to enhance obfuscation while preserving similarity relationships.

Prior research has shown that BFs leak co-occurrence patterns that can be exploited by frequency-based attacks. TMH mitigates some of these weaknesses but still leaks statistical dependencies between hashed values, allowing an attacker to infer recurring n-gram patterns. TSH seeks to combine the efficiency of BFs with additional non-linearity. However, its chained hashing process continues to reveal exploitable patterns. All three encoding schemes remain susceptible to graph-based attacks that leverage the structural properties of similarity graphs derived from encoded data.

Nevertheless, the GMA alone can only recover identities that appear in both D_e and the auxiliary plaintext dataset D_p, leaving non-overlapping records unmapped. The central idea behind the DEA is to leverage the subset of records re-identified by the GMA as labeled examples that link encoded representations to plaintext n-grams. Effectively exploiting this labeled subset with supervised models forms the core of the DEA pipeline and enables inference beyond the original overlap.

2.2 Baseline Frequency-Based Guesser

Before introducing the DEA, we establish a simple, informed baseline that the DEA must outperform. The baseline predicts, for each record, the k most frequent overlapping n-grams (with k set to the dataset specific average record length minus one). Although naive, this frequency-based guesser yields non-trivial performance: $F1 \approx 0.23$ on the `fakename` and `euro_person` datasets, and $F1 \approx 0.29$ on the more heterogeneous `titanic_full`.

This strategy is intentionally simplistic yet plausible. An attacker with access to the global n-gram distribution can perform this guessing strategy across the dataset without any per-record information. Because personal names and dates yield strongly skewed n-gram frequencies, predicting the most common tokens provides a robust, size independent lower bound for reconstruction.

3 Dataset Extension Attack Methodology

3.1 Attack Pipeline Overview

The DEA orchestrates a pipeline that leverages partial re-identifications into broader privacy compromises. First, the attacker runs the improved GMA implementation by Schäfer et al. to obtain an initial set of confirmed plaintext–encoding pairs. These pairs constitute labeled training data for a multi-label classifier. Each encoded record is represented as a fixed-length vector after preprocessing based on the encoding scheme, while the target labels indicate the presence of n-grams in the plaintext. The trained classifier predicts, for each unseen encoded record, the probability that a given n-gram appears in its plaintext. Model selection and hyperparameter tuning optimize the Dice coefficient of preditions to balance precision and recall for reconstructed n-grams, which are subsequently post-processed to generate candidate plaintext reconstructions.

During reconstruction of a complete identifier, the predicted n-grams are leveraged in three reconstruction strategies. The first, greedy reconstruction, uses a graph-based approach in which nodes represent characters and directed edges correspond to predicted n-grams. The second strategy, fuzzy dictionary matching, compares the set of predicted n-grams against a large lookup table of candidate identifiers and ranks matches using the Dice similarity. The third, semantic reconstruction, applies an Large Language Model-based technique to reconstruct identifiers. Taken together, these strategies enable the attacker to either directly reconstruct plaintext identifiers or produce high-confidence candidate lists that can be cross-referenced against additional sources.

3.2 Hyperparameter Optimization and Model Selection

A key component of the DEA is an extensive hyperparameter optimization designed to ensure optimal performance in different attacker scenarios. For each experiment trial performed the DEA performs 125 trials using the Optuna framework. The search space includes model depth (1–3 layers), hidden dimension sizes (64–4096), dropout regularization (0–0.5), activation functions (`ReLU`, `ELU`, `SELU`, `GELU`), optimizers (`Adam`, `AdamW`, `RMSprop`), learning-rate schedulers, batch sizes, and decision thresholds for n-gram prediction. Each trial trains for up to 20 epochs with early stopping (patience of five epochs, minimum validation-loss improvement of 10^{-4}). The Dice coefficient on a held-out validation set guides the optimization process. After convergence, the best-performing configuration is retrained on the full training data to generate predictions on the test set.

3.3 Experimental Design and Evaluation Protocol

The evaluation comprises 180 unique experiments across three datasets (`fakename` with 1k, 2k, 5k, 10k, and 20k records, `euro_person`, and `titanic_full`), three encoding schemes, overlap ratios between 20% and 80%, and two drop-from strategies for the GMA. The "DropFrom = Eve" configuration models an optimistic scenario in which Eve's auxiliary dataset is a strict subset of Alice's database, thereby maximizing overlap quality. In contrast, the "DropFrom = Both" configuration represents a more realistic deployment, where both parties possess unique records, leading to noisier graph structures and weaker training signals for the GMA. For each experiment, the GMA provides the labeled training set, while the remaining encoded records serve as test data for the DEA. Performance is evaluated in terms of structural prediction quality (precision, recall, F1-score, and Dice coefficient) as well as the downstream perfect re-identification rate obtained after fuzzy and greedy reconstruction. Experiments in which the GMA fails to identify any matches are excluded, as the DEA cannot operate without labeled training data in such cases. To ensure comparability, encoding parameters mirror those used in contemporary PPRL studies.

Hyperparameter optimization reveals distinct architectural preferences for each encoding scheme. For TMH, shallow but wide feedforward networks (1024–2048 hidden units) with moderate dropout (≈ 0.25), `ELU/SELU` activations, and `AdamW` optimization perform best. TSH favors slightly deeper architectures with stronger regularization and cyclic learning-rate schedules, reflecting more complex feature interactions from its two-step hashing. BF encodings perform optimally with balanced activations and `RMSprop` combined with cyclic schedules, indicating sensitivity to adaptive, oscillating learning rates. Across all schemes, small batch sizes (8–16) and long training durations highlight the need to fully exploit the limited labeled data provided by the GMA.

4 Empirical Findings

4.1 Tabulation MinHash

Across all datasets, TMH shows the greatest resilience in low-overlap scenarios but becomes increasingly vulnerable to DEAs as more training data becomes available. On the smallest datasets (`fakename_1k` and `fakename_2k`), F1-scores rise from below 0.2 to approximately 0.73 as the overlap increases from 0.4 to 0.8 under "DropFrom = Eve", yet re-identification remains negligible. A turning point appears at `fakename_5k`, where the DEA surpasses an F1-score of 0.85 and achieves the first measurable re-identification rate (1.2%) at 0.6 overlap. Larger datasets amplify this effect: `fakename_10k` reaches near-perfect structural reconstruction (F1 $\approx$ 0.93) with re-identification rates above 5% for overlaps of 0.6 and higher, while `fakename_20k` peaks at 12% re-identification for 0.8 overlap. The more realistic `euro_person` dataset exhibits similar behavior, with F1-scores exceeding 0.9 and re-identification rates up to 6.85%. Overall, these results highlight a pronounced non-linear relationship between structural reconstruction accuracy and deanonymization success: once F1-scores exceed roughly 0.9, even minor gains can trigger steep increases in the number of re-identified individuals.

4.2 Two-Step Hashing

TSH emerges as the most vulnerable encoding overall. Even on the modest `titanic_full` dataset, F1-scores between 0.56 and 0.83 are observed for overlaps of 0.7–0.9, with consistent re-identifications once the overlap exceeds 0.8. The synthetic `fakename` datasets exhibit even faster leakage: `fakename_2k` already achieves F1 = 0.71 and a 1.3% re-identification rate at 0.6 overlap under "DropFrom = Eve". Scaling to `fakename_20k` under favorable overlaps yields near-perfect reconstruction (F1 $\approx$ 0.99) and peak re-identification rates of 28.75%—the highest across all experiments. The `euro_person` dataset confirms this trend, with structural accuracy around 0.96 and re-identification rates up to 12.5%. The additional hashing layer in TSH fails to sufficiently disrupt the statistical dependencies exploited by the DEA, indicating that the chained representation still leaks predictable co-occurrence patterns. Overall, these findings demonstrate that, despite being proposed as a more secure successor to BFs, TSH becomes the weakest link.

4.3 Bloom Filter Encoding

BFs occupy a middle ground: they are more resilient than TSH but considerably weaker than TMH once dataset size and overlap increase. On `fakename_1k`, DEA performance remains close to the frequency baseline, with F1-scores below 0.3 and no successful re-identifications. At `fakename_5k`, however, the DEA surpasses the critical F1 = 0.8 threshold at 0.6 overlap under "DropFrom = Eve", leading to re-identification rates approaching 4%. Larger instances amplify this effect: `fakename_20k` exhibits average re-identification rates around 9%, with peaks near 15% at 0.8 overlap. The `euro_person` dataset follows a similar trajectory, reaching F1 = 0.91 and up to 6.4% re-identification in high-overlap settings. These results indicate that, while BFs degrade gracefully under moderate overlap, their vulnerability increases sharply once sufficient labeled samples enable the DEA to generalize structural patterns across encodings. This finding challenges the common perception of BFs as a "best practice" solution for PPRL and instead positions them as a high-risk choice in environments where attackers can obtain even limited auxiliary overlap data.

4.4 Aggregate Trends and Correlations

Aggregating results across all experiments reveals several consistent patterns. First, overlap size is the primary determinant of DEA success: every encoding shows a monotonic increase in both F1-score and re-identification rate as overlap grows, with the steepest gains occurring between 0.4 and 0.6. Second, dataset size amplifies leakage, as larger datasets provide richer training signals and more homogeneous feature distributions, enabling the neural model to generalize more effectively. Third, the drop-from strategy influences outcomes mainly at low overlaps; once overlap exceeds 0.6, the gap between "DropFrom = Eve" and "DropFrom = Both" narrows considerably, indicating that realistic auxiliary noise offers little protection. Finally, structural accuracy and re-identification follow a sigmoidal relationship: below F1 = 0.7, re-identification remains negligible; between 0.7 and 0.9, leakage rises gradually; and beyond 0.9, small F1 improvements can double or triple the number of compromised records. These dynamics highlight practical thresholds at which defensive monitoring and countermeasures become critical.

5 Implications and Recommendations

The evaluation demonstrates that all three encoding schemes become vulnerable to the DEA once an attacker obtains even a modest training set from the GMA. The resulting privacy implications are twofold. First, the long-standing assumption that non-overlapping records remain safe no longer holds: DEAs can reconstruct n-gram structures with high fidelity and translate them into identifiable plaintext strings. Second, the accessibility of modern hyperparameter optimization frameworks and neural toolkits substantially lowers the barrier for adversaries. Training hundreds of models across 180 scenarios remained computationally feasible within academic resources, implying that well-funded attackers could easily scale and automate such attacks.

To mitigate these risks, several countermeasures are discussed. Increasing the entropy of encoding outputs through diffusion or salting can disrupt the deterministic mappings exploited by the DEA. Integrating cryptographic methods such as secure multi-party computation or homomorphic encryption can decouple similarity computation from direct exposure of encodings, albeit at increased computational cost. However, the findings suggest that incremental adjustments to existing encodings are unlikely to provide meaningful protection; a shift toward cryptographically rigorous linkage protocols may be required for high-stakes applications.

6 Conclusion and Outlook

This summary consolidates the thesis' main contributions: the formalization of the DEA threat model, the implementation of a comprehensive neural attack pipeline, and the empirical demonstration of vulnerabilities in widely adopted PPRL encodings under realistic conditions. The results show that TSH is the most susceptible scheme, BFs offer only moderate resistance, and TMH, while comparatively resilient, remains vulnerable when sufficient overlap and training data are available. Across 180 experiments, the DEA achieves peak re-identification rates of 28.75% and maintains average structural F1-scores above 0.6, substantially outperforming naive frequency-based baselines. These findings demonstrate that privacy assurances relying solely on obscurity or limited overlap are no longer sustainable.

Future research should advance along three directions. First, defensive measures should establish principled bounds on information leakage from similarity-preserving encodings, for instance by integrating differential privacy or cryptographic commitments. Second, future attackers may extend the DEA through generative modeling, adversarial learning, or transfer learning to generalize across domains, emphasizing the need for proactive mitigation strategies. Third, policy and governance frameworks should mandate formal risk assessments before deploying PPRL systems in sensitive domains. By exposing the dataset-extension threat, this work provides both a cautionary perspective and a foundation for developing the next generation of privacy-preserving linkage protocols.